

Herramientas de transcripción asistida colaborativa de censos históricos

1.- Introducción y Objetivos

El gran volumen de documentos almacenados en archivos históricos son un patrimonio de gran relevancia para el estudio y evolución de las sociedades que contribuyen a la preservación de la memoria histórica. En la era digital, las bibliotecas y archivos han dedicado un gran esfuerzo a digitalizar de forma masiva su documentación de carácter histórico. De este modo se asegura su preservación, pero a su vez, se abren nuevos retos sobre el acceso y valorización de los documentos digitales a través de la extracción, indexación y vinculación de sus contenidos mediante herramientas informáticas.

Las humanidades digitales son un área emergente e interdisciplinar en la que convergen las humanidades y la informática. A través del proyecto *"EINES"*, financiado por *La Obra Social La Caixa*, investigadores de las áreas de demografía y ciencias de la computación se unen para desarrollar instrumentos y procedimientos que faciliten la informatización masiva de las fuentes demográficas como los padrones. El interés de los padrones reside en que son una fuente que a lo largo del siglo XIX se generalizó a la inmensa mayoría de municipios españoles, y que de forma similar, existen en muchos otros países, siendo por tanto, un proyecto escalable a nivel europeo.

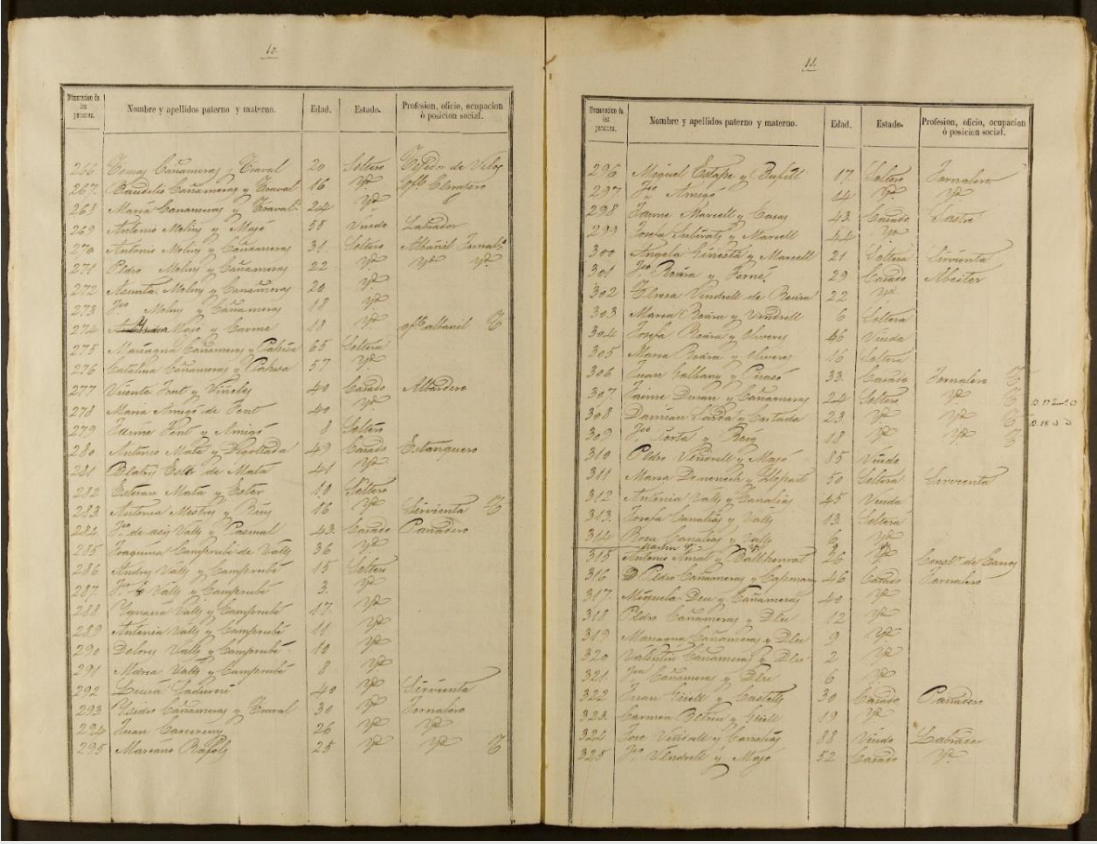
Los objetivos del proyecto son: construir bases de datos de uso público, mejorar el acceso y consulta de los documentos de archivos y construir herramientas de análisis de los datos.

En concreto, en este proyecto pretendemos:

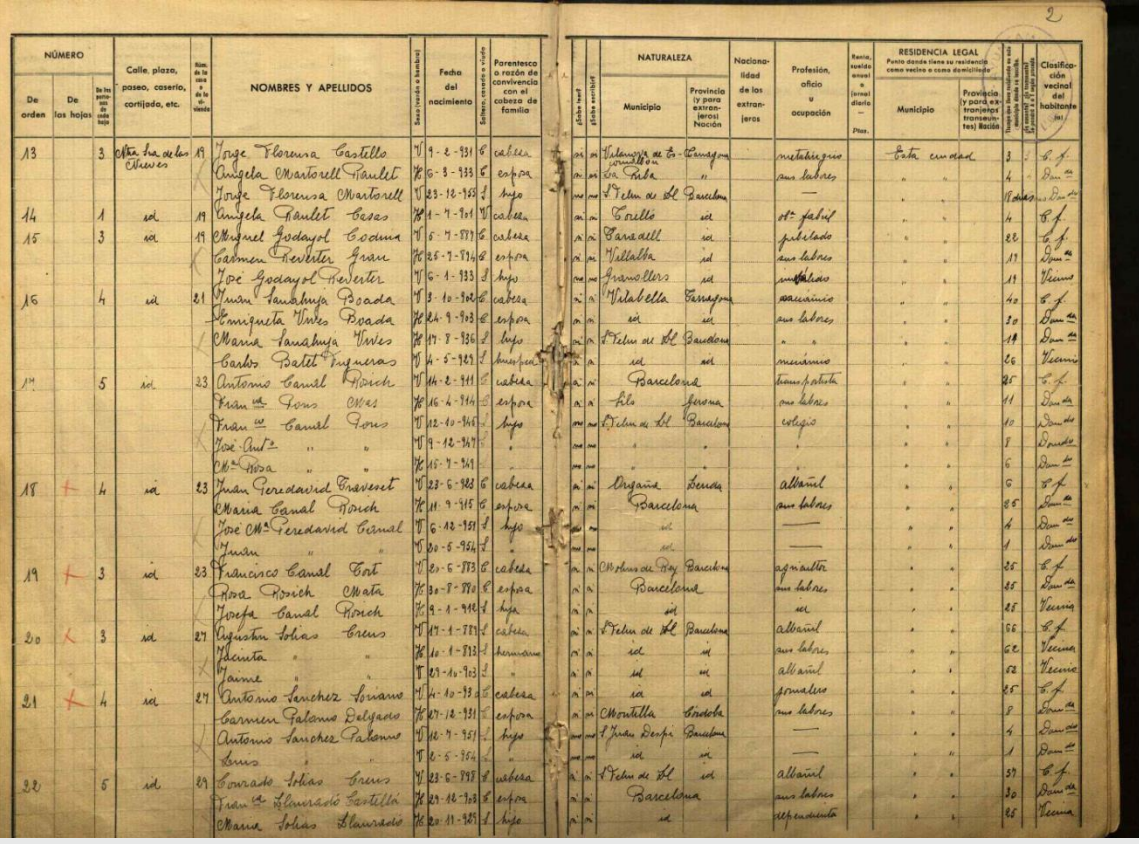
1. Experimentar en la **transcripción semi-asistida** mediante dos procedimientos alternativos:
 - Reconocer los términos más frecuentes (top ten)
 - Arrastre cronológico de la información (para facilitar la transcripción y mostrar la dinámica de hogares).
2. Experimentar en una plataforma de **transcripción colaborativa** (crowdsourcing).
3. Utilización de técnicas de *"record linkage"*: establecer trayectorias individuales a lo largo del tiempo, construcción automática de genealogías, localización espacial de redes familiares.

2.- La colección

- Se han seleccionado los padrones y censos del municipio de *Sant Feliu de Llobregat*, por ser un municipio importante de la época, con una estructura social y ocupacional diversificada. Contaba con 1.579 habitantes en 1787, 2.478 habitantes en 1860, 3.314 habitantes en 1900, y 7.327 habitantes en 1950.
- Se realizaron 19 censos entre 1828 y 1955. Aproximadamente 30.000 páginas (escaneadas a alta calidad).
- Los censos suelen recoger la siguiente información: La población de hecho y la población de derecho; el estatus migratorio; la distribución por sexo, edad y estado civil de la población; la alfabetización o grado de instrucción; la población activa; la fecundidad; las características familiares de la población, y las características de la vivienda.



Censo de 1857

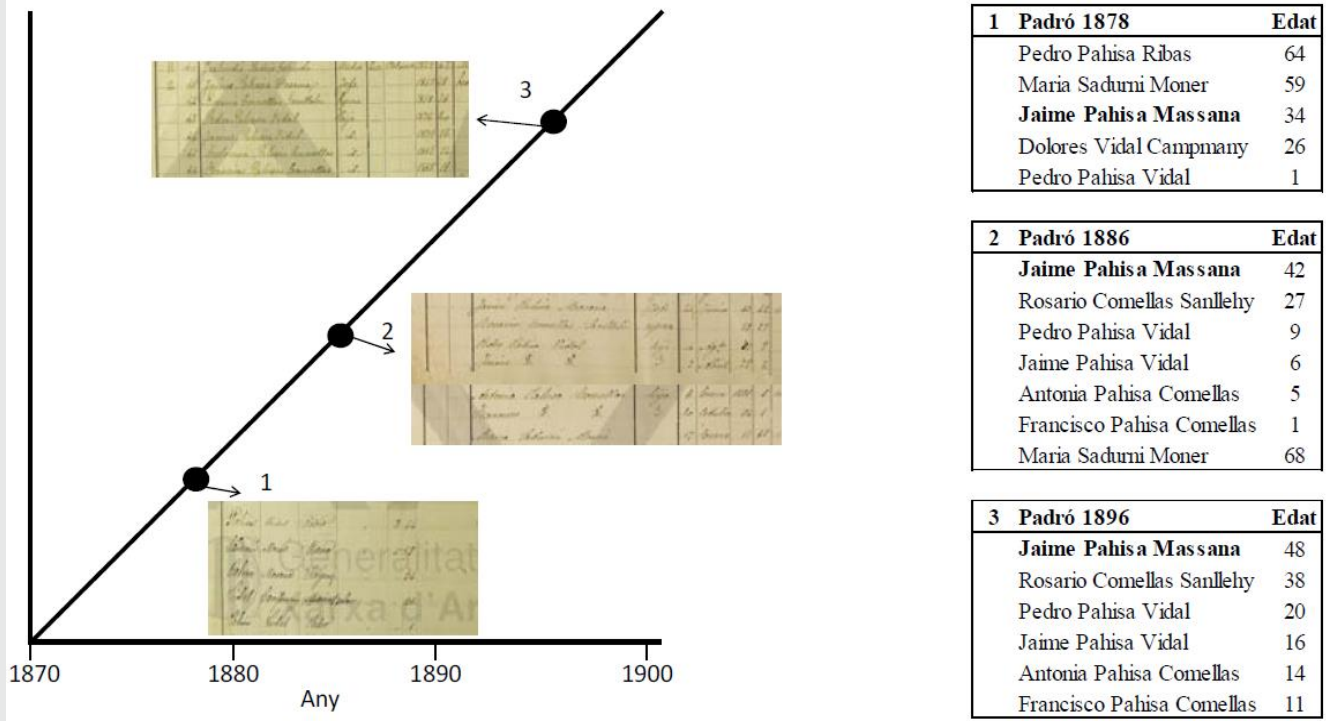


Censo de 1955

4.- La construcción de trayectorias vitales

A través del *"record linkage"* (vinculación nominal de registros) se pretende automatizar la generación de trayectorias individuales y familiares a lo largo del tiempo, así como la localización espacial de redes familiares. El resultado final será una gran red social histórica con multitud de genealogías entrelazadas. Este *"record linkage"* consiste en enlazar por ejemplo padres e hijos o detectar las apariciones del mismo individuo a lo largo de padrones de años sucesivos, etc.

Dado que muchos nombres, apellidos y lugares pueden aparecer con variantes ortográficas, en la plataforma de transcripción se están incorporando varias técnicas de comparación de cadenas de caracteres, como la distancia de *Levenshtein* [3]. Así, para cada registro, el sistema propone enlaces entre individuos que tienen nombres y apellidos muy similares que vivan en el mismo lugar (ej. misma calle). Finalmente, el experto valida las transcripciones y las relaciones propuestas entre individuos.



5.- Conclusiones

La transcripción de los padrones de *Sant Feliu de Llobregat* y la generación de trayectorias de vida no es un fin en sí mismo, sino que se busca que sea el prototipo que permita experimentar para después adaptar las herramientas generadas a otras fuentes documentales o municipios distintos. El objetivo final es poner la información demográfica histórica al servicio de la ciudadanía, creando nuevos productos y servicios a disposición de la sociedad para facilitar el acceso universal a los archivos históricos y que a partir de ésta se pueda generar conocimiento.

Agradecimientos

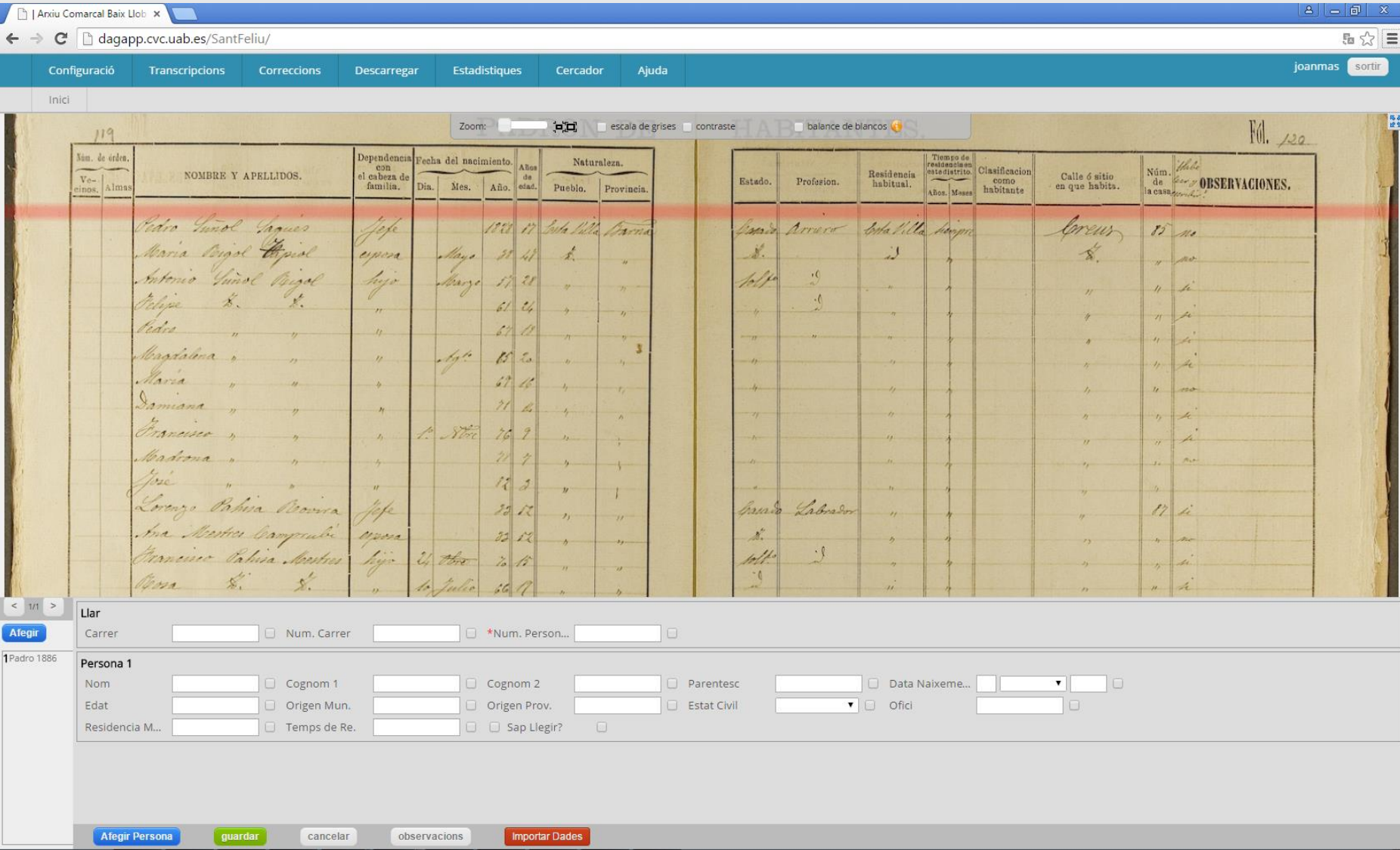
Este proyecto ha contado con el soporte de RecerCaixa, un programa impulsado por la Obra Social "La Caixa", con la colaboración de la ACUP. También ha sido parcialmente financiado por el European Research Council Advanced Grant (ERC-2010-AdG 20100407: 269796-5CoFm) y el Proyecto nacional TIN2012-37475-C02-02. Los autores agradecen especialmente al Arxiu Comarcal del Baix Llobregat por las imágenes y a los transcripores por su labor desinteresada en la transcripción.

3.- Plataforma de transcripción asistida colaborativa

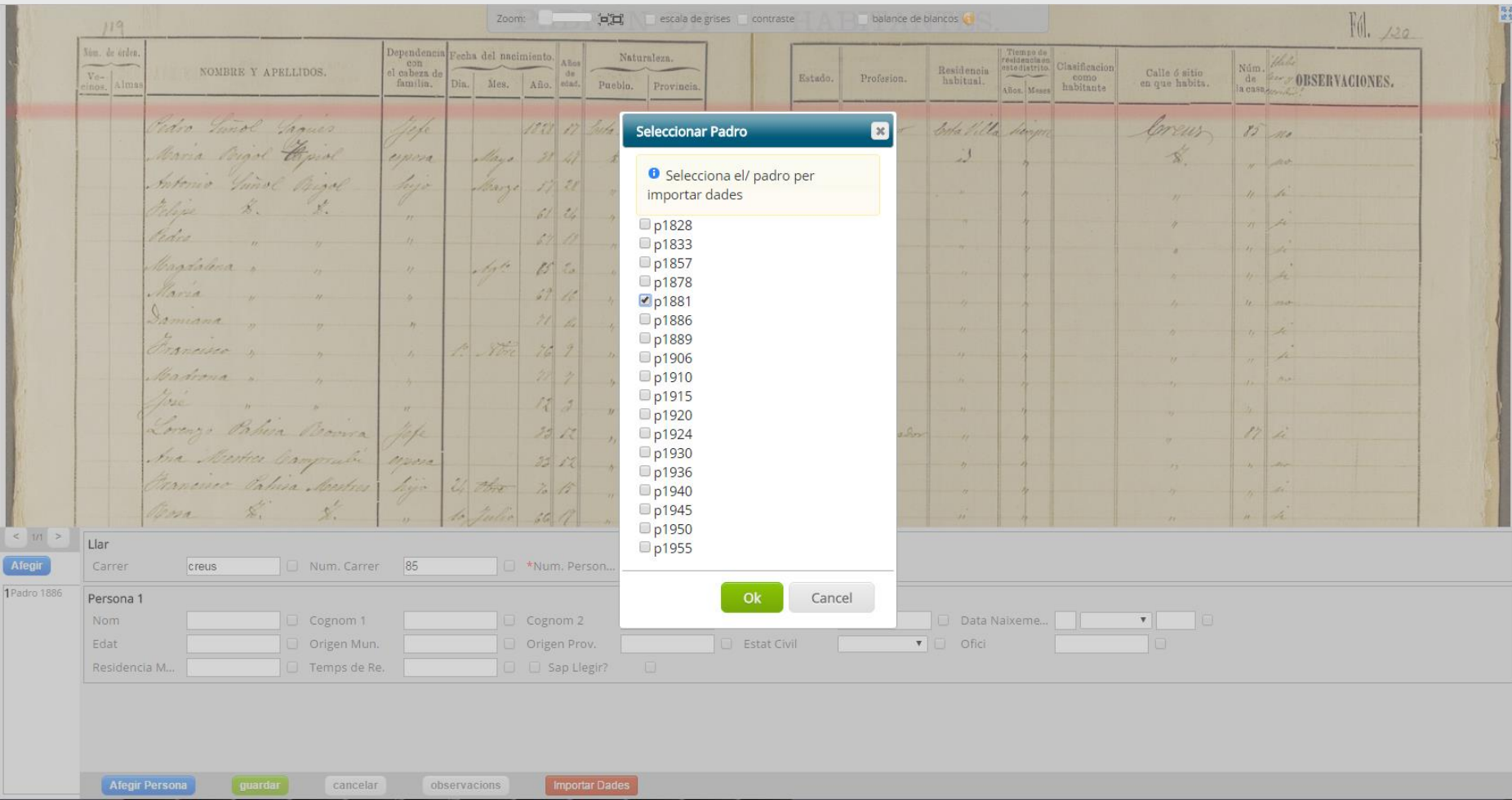
La plataforma de acceso a través de internet permite a los usuarios transcribir de forma simultánea los contenidos de las imágenes del archivo, así como validar las transcripciones por parte de los expertos para asegurar la consistencia de los datos. La plataforma se basa en el paradigma de *"crowdsourcing"*, que ya se ha seguido para transcribir registros matrimoniales [1, 2]. Así, la transcripción se divide en múltiples tareas pequeñas (ej. páginas) que puedan llevar a cabo un grupo numeroso de transcripores.

Transcripción semi-asistida. La plataforma incorporará progresivamente herramientas para reducir el tiempo dedicado a la transcripción. La primera herramienta incorporada es la del **arrastre cronológico**, que consiste en permitir la importación selectiva de datos de un censo anterior ya transcrito. De este modo, cuando el usuario selecciona un documento (paso 1) e introduce la calle y número, el sistema le ofrece importar datos de ese mismo hogar de un censo ya transcrito (paso 2). Entonces, el usuario selecciona qué habitantes de ese hogar desea importar (paso 3). Como resultado, sólo tiene que actualizar algún dato (ej. edades), y si es necesario, añadir algún habitante nuevo (paso 4).

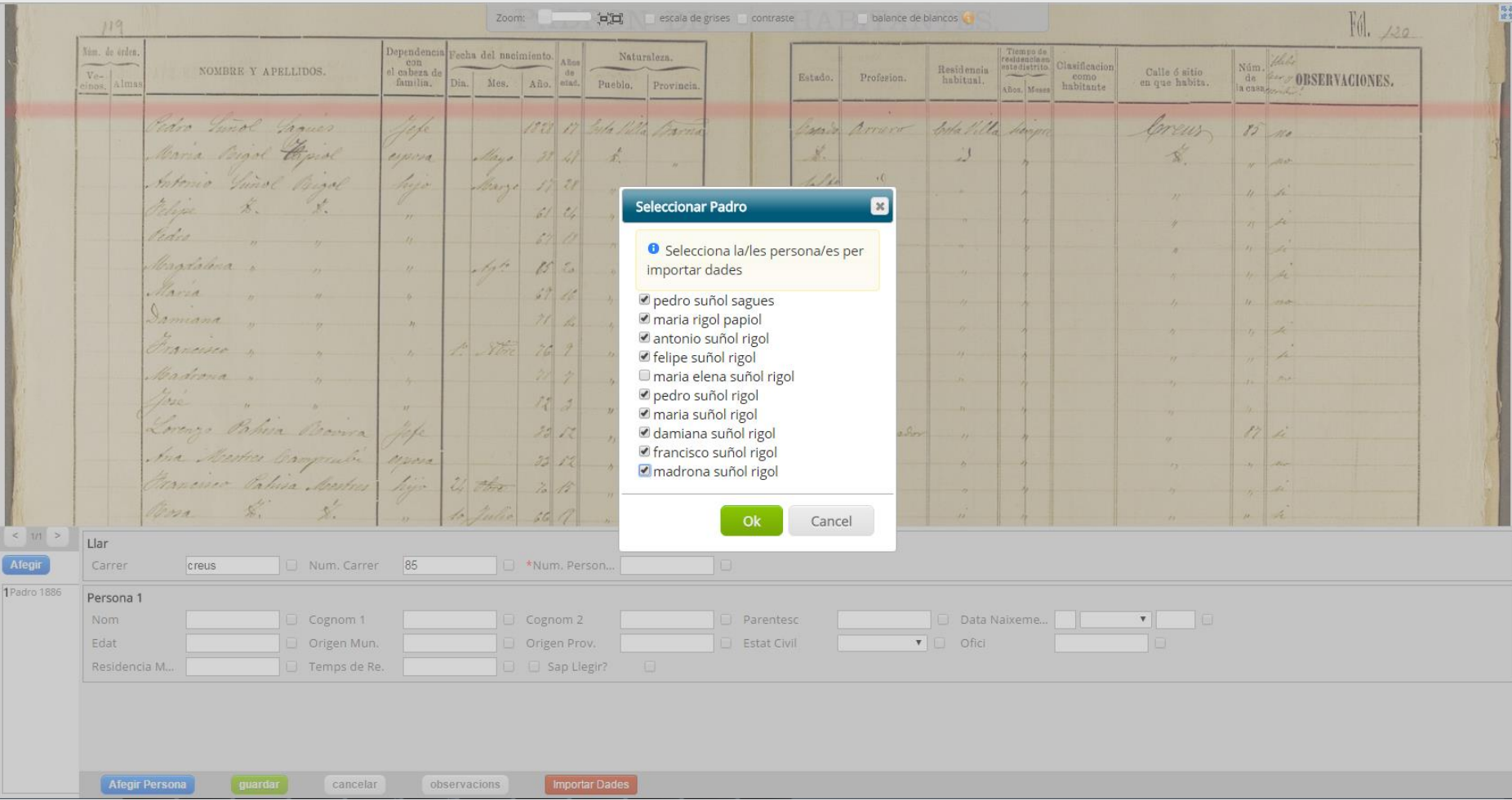
Paso 1



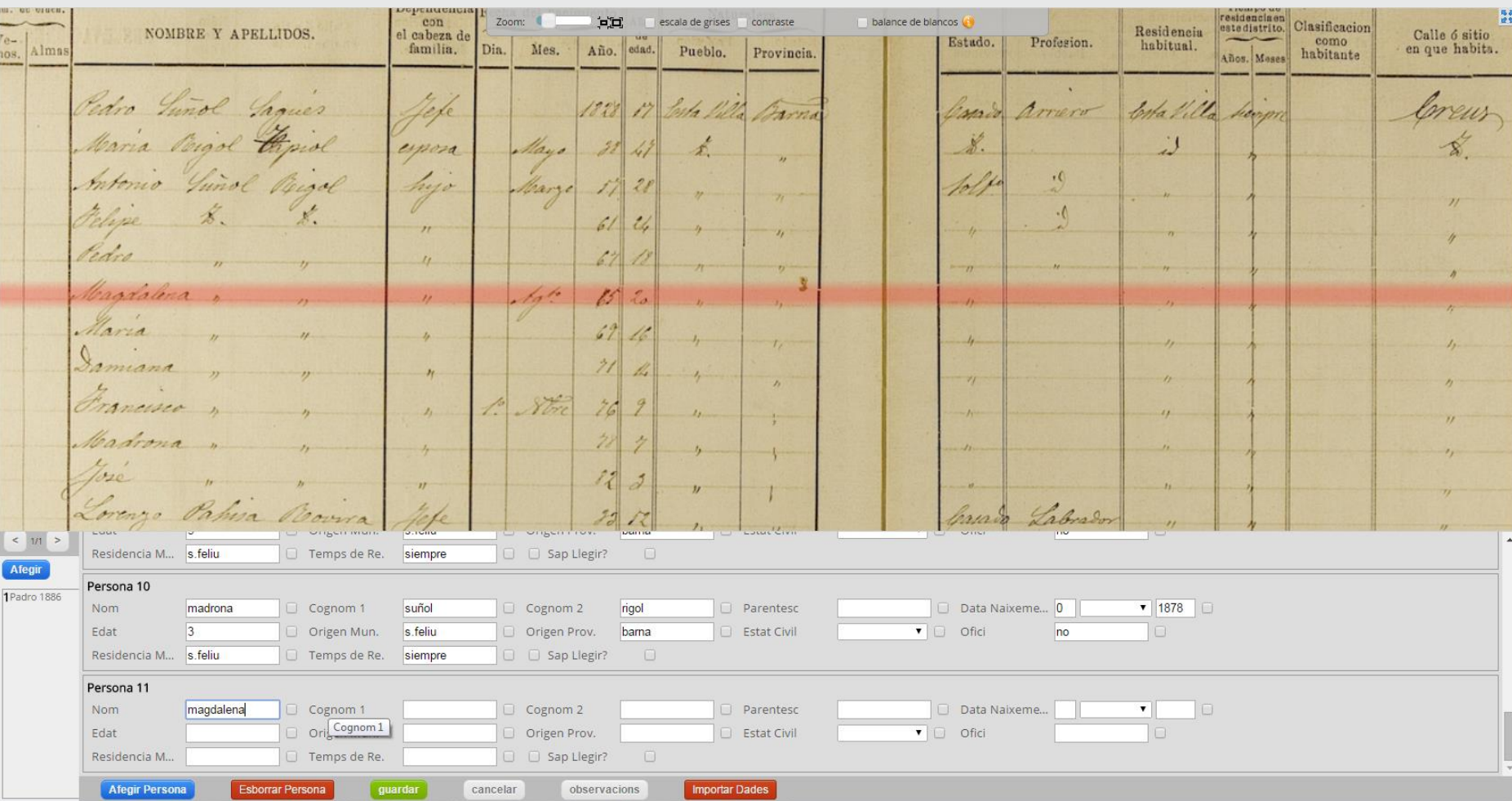
Paso 2



Paso 3



Paso 4



Referencias

- 1) G.Thorvaldsen, J.M.Pujades-Mora, T.Andersen, L.Eikvil, J.Lladós, A.Fornés, A.Cabré. "A Tale of two Transcriptions". Historical Life Course Studies, Volume 2, pp. 1-19, (e-ISSN-2352-6343), 2015.
- 2) A.Fornés, J.Lladós, J.Mas, J.M.Pujades, A.Cabré. "A Bimodal Crowdsourcing Platform for Demographic Historical Manuscripts". Digital Access to Textual Cultural Heritage Conference (DATECH), pp. 103-108, 2014.
- 3) G. Navarro. A guided tour to approximate string matching. ACM Computing Surveys, 33(1), pp.31-88, 2001.